

Prompt, Generate, Then Cache: Cascade of Foundation Models Makes Strong Few-Shot Learners

Year: 2023 | Citations: 216 | Authors: Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng

Abstract

Visual recognition in low-data regimes requires deep neural networks to learn generalized representations from limited training samples. Recently, CLIP-based methods have shown promising few-shot performance benefited from the contrastive language-image pre-training. We then question, if the more diverse pre-training knowledge can be cascaded to further assist few-shot representation learning. In this paper, we propose CaFo, a Cascade of Foundation models that incorporates diverse prior knowledge of various pretraining paradigms for better few-shot learning. Our CaFo incorporates CLIP's language-contrastive knowledge, DINO's vision-contrastive knowledge, DALL-E's vision-generative knowledge, and GPT-3's language-generative knowledge. Specifically, CaFo works by 'Prompt, Generate, then Cache'. Firstly, we leverage GPT-3 to produce textual inputs for prompting CLIP with rich downstream linguistic semantics. Then, we generate synthetic images via DALL-E to expand the few-shot training data without any manpower. At last, we introduce a learnable cache model to adaptively blend the predictions from CLIP and DINO. By such collaboration, CaFo can fully unleash the potential of different pre-training methods and unify them to perform state-of-the-art for few-shot classification. Code is available at <https://github.com/ZrrSkywalker/CaFo>.